Name : JasRirat Singh
class : CSE-1 , 8th Semester
Roll No. : 05413202717
Subject : WeB Intelligance and Big Data
Paper Code : ETCS -458

Aim : How to handle missing data in pandas using fillna , interpolate and drop na methods

Theory :
One of the most introductory Big data interview questions asked during interviews, the answer to this is fairly straightforward

Big data is defined as a collection of large and complex unstructured data sets from where insights are derived from Data Analysis using opn-source tools like hadoop

The five Vs of Big Data are -

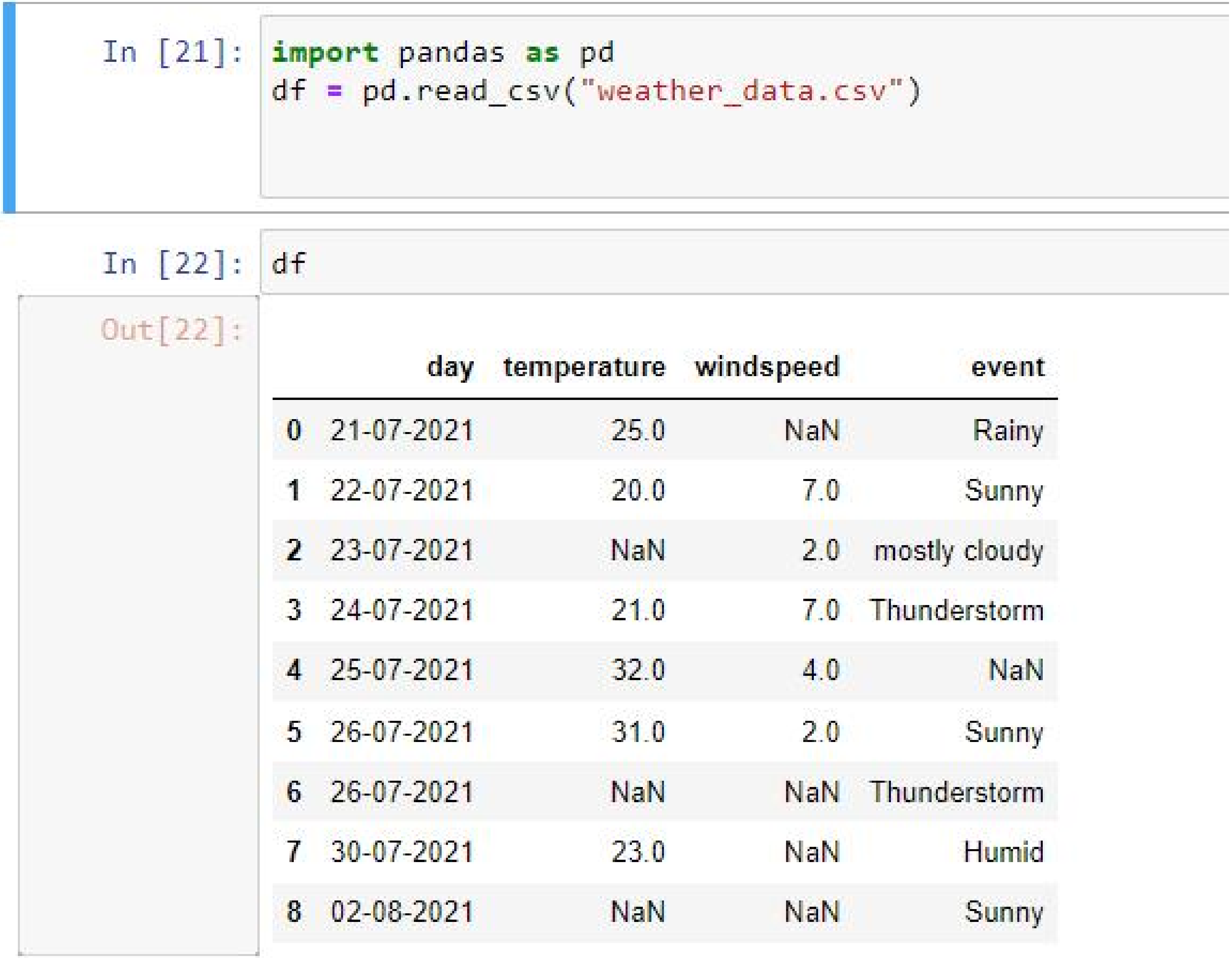
Volume - Amount of data in Petabytes and Exabytes
Variety - includes formats like videos , audios sources textual data etc
Velocity - Everyday data growth which includes conversation in forums blogs, Social Media posts etc

Veracity - Degree of accuracy of data variable

Value - Deriving insights from collected data to
achieve business millstones
and new heights

Code:

```
import pandas as pd

df = pd.read_csv("weather_data.csv")
df

df = pd.read_csv("weather_data.csv, parse_dates=['day'])
type(df.day[0])
df

df.set_index('day', inplace=True)
df

new_df = df.fillna(0)
new_df

new_df = df.fillna({
        'temperature': 0,
        'windspeed': 0,
        'event': 'No Event'
new_df
```



In [21]:

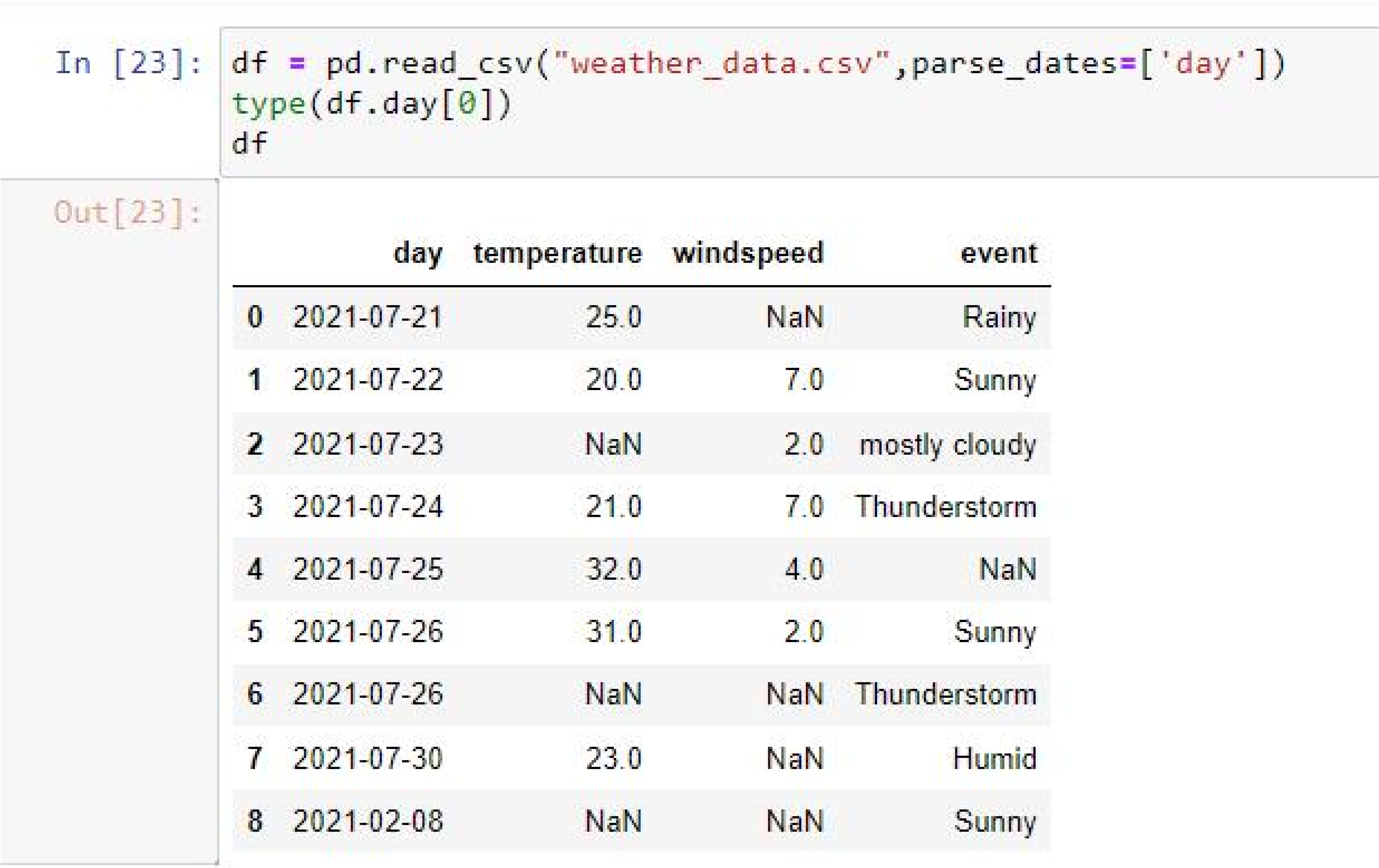
```
import pandas as pd
df = pd.read_csv("weather_data.csv")
```

In [22]:

```
df
```

Out[22]:

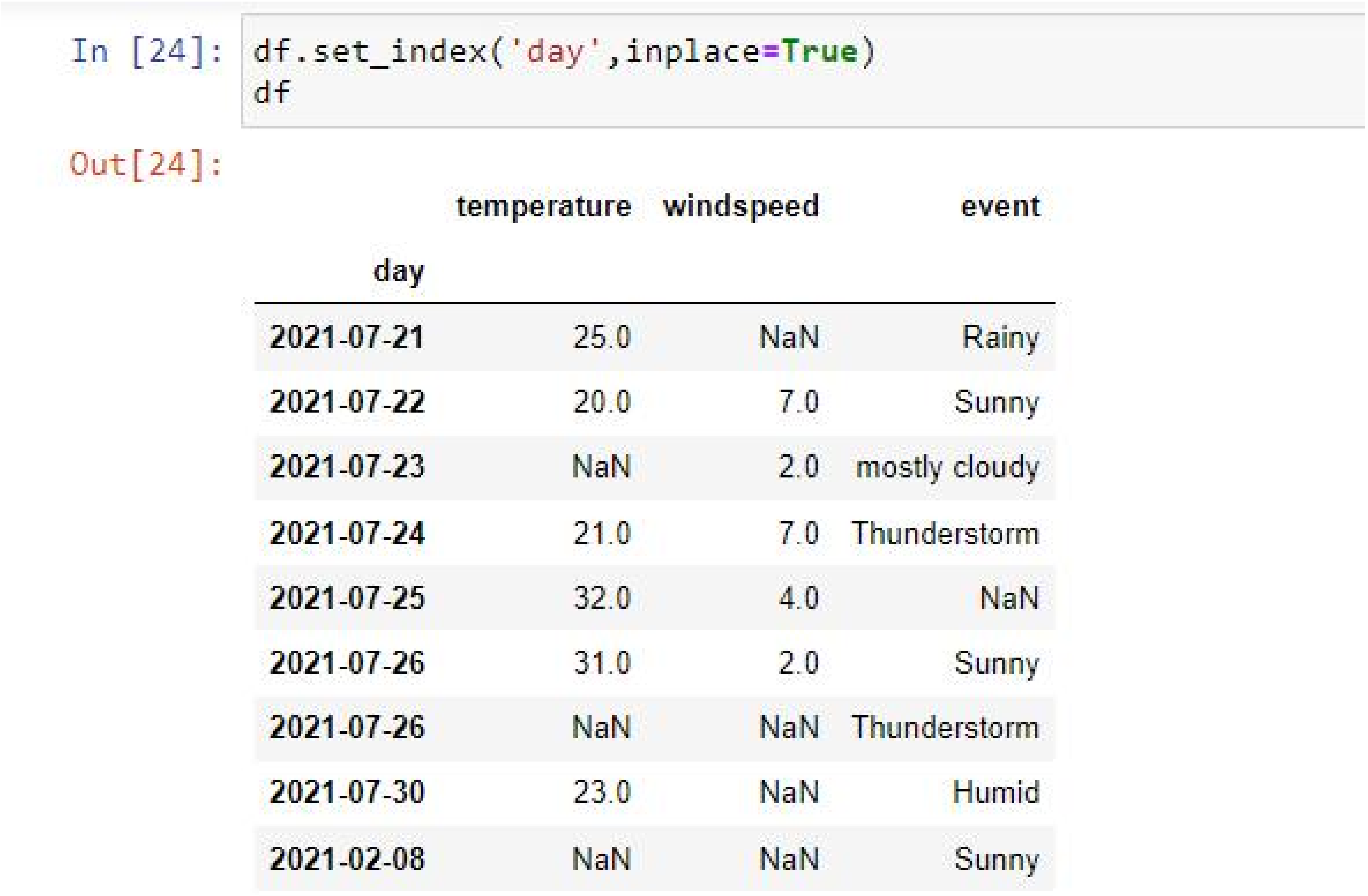
	day	temperature	windspeed	event
0	21-07-2021	25.0	NaN	Rainy
1	22-07-2021	20.0	7.0	Sunny
2	23-07-2021	NaN	2.0	mostly cloudy
3	24-07-2021	21.0	7.0	Thunderstorm
4	25-07-2021	32.0	4.0	NaN
5	26-07-2021	31.0	2.0	Sunny
6	26-07-2021	NaN	NaN	Thunderstorm
7	30-07-2021	23.0	NaN	Humid
8	02-08-2021	NaN	NaN	Sunny



```
In [23]: df = pd.read_csv("weather_data.csv",parse_dates=['day'])
         type(df.day[0])
         df
```

Out[23]:

	day	temperature	windspeed	event
0	2021-07-21	25.0	NaN	Rainy
1	2021-07-22	20.0	7.0	Sunny
2	2021-07-23	NaN	2.0	mostly cloudy
3	2021-07-24	21.0	7.0	Thunderstorm
4	2021-07-25	32.0	4.0	NaN
5	2021-07-26	31.0	2.0	Sunny
6	2021-07-26	NaN	NaN	Thunderstorm
7	2021-07-30	23.0	NaN	Humid
8	2021-02-08	NaN	NaN	Sunny

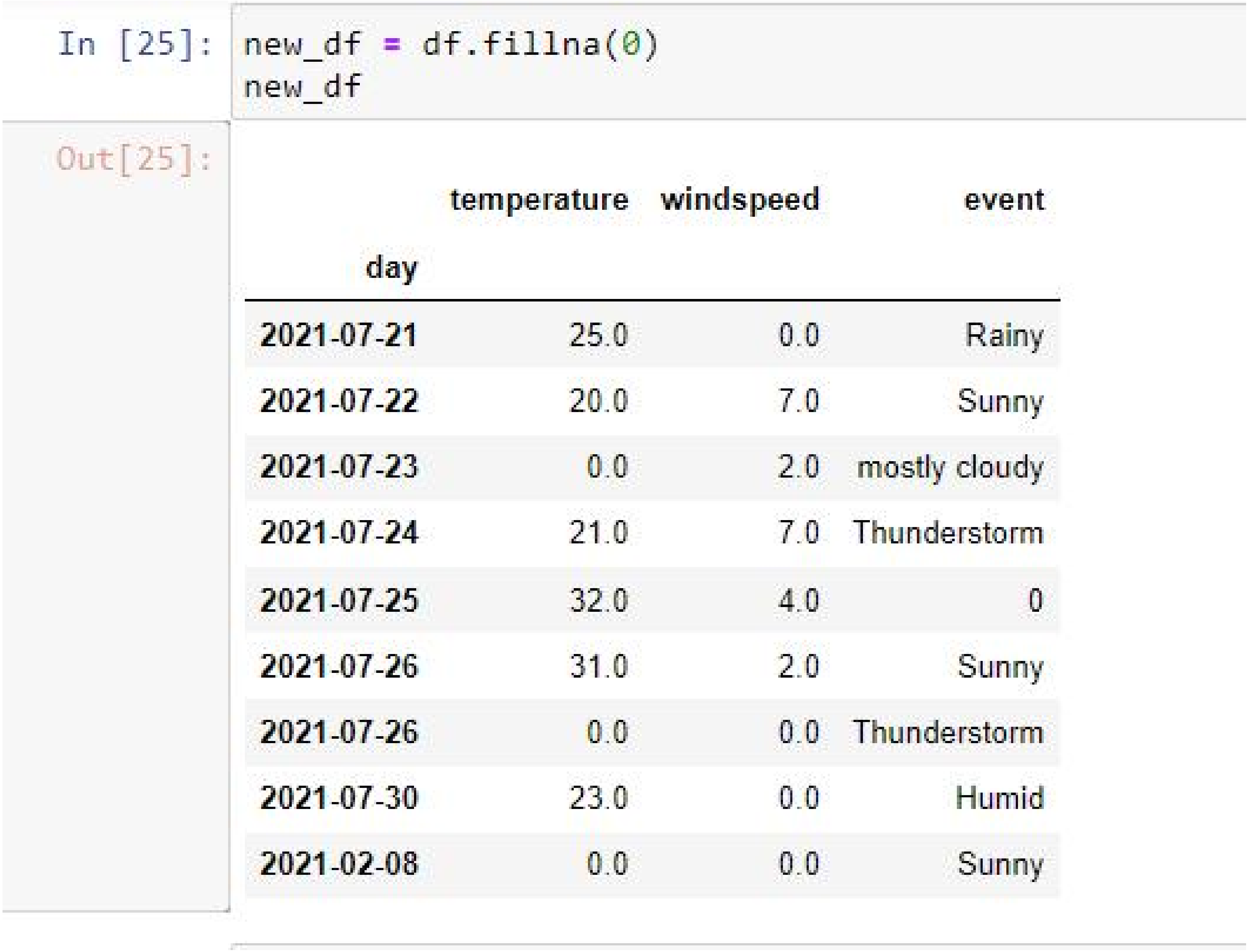


In [24]:

```
df.set_index('day',inplace=True)
df
```

Out[24]:

	temperature	windspeed	event
day			
2021-07-21	25.0	NaN	Rainy
2021-07-22	20.0	7.0	Sunny
2021-07-23	NaN	2.0	mostly cloudy
2021-07-24	21.0	7.0	Thunderstorm
2021-07-25	32.0	4.0	NaN
2021-07-26	31.0	2.0	Sunny
2021-07-26	NaN	NaN	Thunderstorm
2021-07-30	23.0	NaN	Humid
2021-02-08	NaN	NaN	Sunny

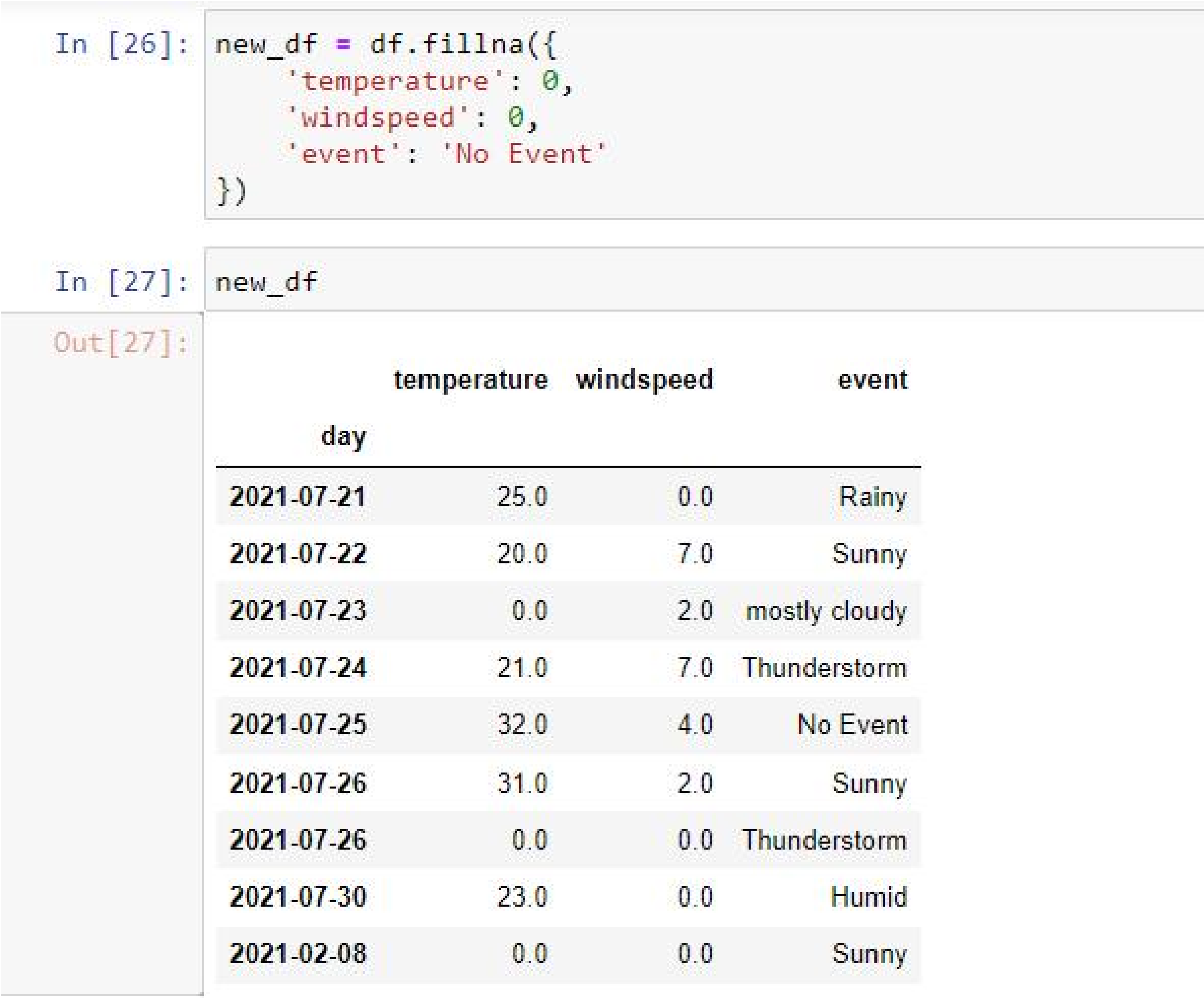


In [25]:

```
new_df = df.fillna(0)
new_df
```

Out[25]:

	temperature	windspeed	event
day			
2021-07-21	25.0	0.0	Rainy
2021-07-22	20.0	7.0	Sunny
2021-07-23	0.0	2.0	mostly cloudy
2021-07-24	21.0	7.0	Thunderstorm
2021-07-25	32.0	4.0	0
2021-07-26	31.0	2.0	Sunny
2021-07-26	0.0	0.0	Thunderstorm
2021-07-30	23.0	0.0	Humid
2021-02-08	0.0	0.0	Sunny



In [26]:

```
new_df = df.fillna({
    'temperature': 0,
    'windspeed': 0,
    'event': 'No Event'
})
```

In [27]:

```
new_df
```

Out[27]:

	temperature	windspeed	event
day			
2021-07-21	25.0	0.0	Rainy
2021-07-22	20.0	7.0	Sunny
2021-07-23	0.0	2.0	mostly cloudy
2021-07-24	21.0	7.0	Thunderstorm
2021-07-25	32.0	4.0	No Event
2021-07-26	31.0	2.0	Sunny
2021-07-26	0.0	0.0	Thunderstorm
2021-07-30	23.0	0.0	Humid
2021-02-08	0.0	0.0	Sunny